Homework 2

Introduction to Data Visualization

FLINT

- Please make sure that you upload the .R file to Canvas and provide a link to your GitHub repository where you need to upload your .R script.
- Please also submit a pdf file which includes your code snippets and their outputs (just for Canvas). You don't have to output the whole data frame. You can consider using **head** function.
- Please do not just submit your answers. Your submission should demonstrate the code and the output for every question.
- No late submissions will be accepted **unless you have an excuse**.
- You may need to show your answers to me during the class.

1. Run the following lines and study how they work. Then state what they do and output for us. (20 Points)

```
df1=data.frame(Name=c('James','Paul','Richards','Marico','Samantha','Ravi','Raghu',
     'Richards','George','Ema','Samantha','Catherine'),

     State=c('Alaska','California','Texas','North Carolina','California','Texas',
     'Alaska','Texas','North Carolina','Alaska','California','Texas'),

          Sales=c(14,24,31,12,13,7,9,31,18,16,18,14))

aggregate(df1$Sales, by=list(df1$State), FUN=sum)


library(dplyr)
df1 %>% group_by(State) %>% summarise(sum_sales = sum(Sales))
```

2. Use **R** to read the **WorldCupMatches.csv** from the DATA folder on Google Drive. Then perform the followings (48 points):
   (a) Find the size of the data frame. How many rows, how many columns?
   (b) Use summary function to report the statistical summary of your data.
   (c) Find how many unique locations olympics were held at.
   (d) Find the average attendance.
   (e) For each Home Team, what is the total number of goals scored? (Hint: Please refer to question 1)
   (f) What is the average number of attendees for each year? Is there a trend or pattern in the data in that sense?
3. Use **R** to read the **metabolites.csv** from the DATA folder on Google Drive. Then perform the followings (32 points):
   (a) Find how many Alzheimers patients there are in the data set. (Hint: Please refer to question 1)
   (b) Determine the number of missing values for each column. (Hint: is.na( ))
   (c) Remove the rows which has missing value for the **Dopamine** column and assign the result to a new data frame. (Hint: is.na( ))
   (d) In the new data frame, replace the missing values in the **c4-OH-Pro** column with the median value of the same column. (Hint: there is median( ) function.)
   (e) (Optional) Drop columns which have more than 25% missing values. (Hint: when you slice your data frame, you can use -c(.., ..., ...) where ... represent one column name)